Available online at www.sciencedirect.com

SCIENCE DIRECT®

Vision Research 46 (2006) 2535–2545

Vision Research

www.elsevier.com/locate/visres

# Fixations in natural scenes: Interaction of image structure and image content

Christoph Kayser *, Kristina J. Nielsen, Nikos K. Logothetis

*Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany*

Received 30 September 2005; received in revised form 23 January 2006

**Abstract**

Explorative eye movements specifically target some parts of a scene while ignoring others. Here, we investigate how local image structure—defined by spatial frequency contrast—and informative image content—defined by higher order image statistics—are weighted for the selection of fixation points. We measured eye movements of macaque monkeys freely viewing a set of natural and manipulated images outside a particular task. To probe the effect of scene content, we locally introduced patches of pink noise into natural images, and to probe the interaction with image structure, we altered the contrast of the noise. We found that fixations specifically targeted the natural image parts and spared the uninformative noise patches. However, both increasing and decreasing the contrast of the noise attracted more fixations, and, in the extreme cases, compensated the effect of missing content. Introducing patches from another natural image led to similar results. In all paradigms tested, the interaction between scene structure and informative scene content was the same in any of the first six fixations on an image, demonstrating that the weighting of these factors is constant during viewing of an image. These results question theories, which suggest that initial fixations are driven by stimulus structure whereas later fixations are determined by informative scene content.


*Keywords:* Eye movements; Luminance–contrast; Natural image; Macaque monkey; Image semantics; Pink noise

## 1. Introduction

Every day vision is an active process, in which only a small portion of our visual environment is selected for thorough analysis. Saccadic eye movements direct our gaze to specific locations in a scene and bias our perception towards specific features. Already the first systematic studies of human eye movements demonstrated that fixations specifically target locations that are 'informative or useful for perception' (Buswell, 1935; Yarbus, 1967). Recent results extended these findings and suggest that fixations are determined by either the salient structure of local visual features or by the image content.

Fixations in natural scenes specifically target locations characterized by distinct physical image properties such as high luminance–contrast and high edge density (Einhäuser & König, 2003; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Mannan, Ruddock, & Wooding, 1996, 1997; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005). Exploiting such rudimentary image properties, models were proposed that successfully predict human fixation patterns (Itti & Koch, 1998, 2000, 2001; Parkhurst, Law, & Niebur, 2002; Parkhurst & Niebur, 2004). Importantly, several studies reported that the effect of saliency—i.e., the conspicuity of the visual features—was strongest for the initial fixation on an image, but that it decreased over time, usually during the first four to six fixations (Itti, 2005a; Parkhurst et al., 2002).

Other studies revealed an effect of informative image content on fixations. Fixations specifically target locations rated as informative by other humans (Antes, 1974; Mackworth & Morandi, 1967; Yarbus, 1967) and target

* Corresponding author.
*E-mail address:* kayser@tuebingen.mpg.de (C. Kayser).

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.02.003

locations 'inconsistent with the global gist of a scene' (Loftus & Mackworth, 1978). While these effects are robustly found on average, some studies did not observe these during the initial (three or four) fixations (De Graef, Christiaens, & d'Ydewalle, 1990; Henderson, Weeks, & Hollingworth, 1999; Hollingworth & Henderson, 1998), but see (Antes, 1974). Hence, in agreement with the above, this led to the suggestion that the placement of initial fixations is determined by physical image structure, while the influence of informative scene content becomes prominent during later fixations (Henderson & Hollingworth, 1999).

While these results suggest that the weighting of informative content with spatial image structure is a prominent factor for determining scan paths, there are several unresolved issues. First, recent studies could not replicate the temporal effect (Einhäuser, Kruse, Hoffmann, & König, 2006; Tatler et al., 2005). Second, many of the above studies were conducted using highly simplified images or gross manipulations, leaving open the trade-off between salient scene structure and content in natural settings. And third, studies with human subjects often employ a specific task, but seldom have access to naïve subjects. The present study was designed to overcome these limitations.

We analyzed the fixations of naïve macaque monkeys on a large set of manipulated natural images that were presented for free-viewing without imposing a particular task. Different manipulations of image structure, which is here defined by the spatial frequency contrast, and of informative image content, which is defined by the higher order image statistics were used. A first manipulation was designed to preserve the local image structure while altering the image content by introducing uninformative noise patches. This manipulation preserved the second order statistics of an image, but locally introduced a randomized higher order structure. A second manipulation introduced a dissociation between scene content and scene structure, by altering the luminance–contrast of the local noise patches. Finally, in a third manipulation, the local information content about an image was altered by blending patches from one natural image into another. This manipulation preserves the overall kind of image statistics, by using only natural images, but changes the relation between the global image and the local manipulation. Our results demonstrate how local scene structure (spatial frequency contrast) and informative content interact to determine fixated locations, how contrast can compensate for missing image content and clearly show that this effect is the same during the first six fixations on an image.

## 2. Materials and methods

### 2.1. Visual stimuli

A set of 40 natural images containing different landscapes and animals was used. Only three images contained prominent man made artifacts (city views). To construct manipulated images, we first generated a set of pink-noise images. Pink noise has the same frequency (Fourier) spectrum as the corresponding natural image but has a random higher order statistical structure. These images were obtained by Fourier transforming the original image, replacing the phases with random values between 0 and 2pi, and applying the inverse Fourier transform. This manipulation preserves the Fourier amplitudes and hence the spatial frequency composition of an image, but it destroys the higher order structure defined by the phases of individual frequencies. As a result, the pink-noise images appear as 'cloud-like' images without particular objects as they would occur in a natural image. All natural and noise images were calibrated to have same mean luminance and root-mean-squared contrast. These natural and full pink-noise images were used initially to detect possible gross differences in fixation patterns.

For the first paradigm, the manipulated images consisted of natural images in which local patches of pink noise were introduced (see Fig. 1A for examples). To locally blend a patch of noise into the original image, a mask was created. This mask consisted of 20 Gaussian kernels, distributed pseudo-randomly across the image (see Fig. 1B). Each kernel had a standard deviation ($\sigma$) of 1.2 deg and their overlap was limited to less than 15%. The values of this mask defined the blending with a value of zero defining a purely natural image and a value of one defining a pure noise image:

$$\text{modified\_image} = (1 - \text{mask}) * \text{natural\_image} + \text{mask} * \text{noise\_image}. \quad (1)$$

A separate mask was created for each image. Importantly, the mean luminance and root-mean-squared contrast for the noise patch were adjusted to the values of the natural image in this area, which were estimated from the inner 15% of the area covered by the kernel. The resulting images are globally similar to the original natural images but locally contain uninformative noisy regions. Importantly, outside the modified regions, these images are identical to the original images. Animal M1 was presented five such sets consisting of 40 manipulated images each (200 images in total), animal M2 was presented three sets (120 images), and animals M3 and M4 viewed two sets (80 images).

A second set of manipulated images was used in the second paradigm (Fig. 3A). This was obtained similarly as in the first paradigm, but the contrast of the blended noise patch was manipulated. After adjusting luminance and contrast of the blended patch to that of the original image, the contrast was scaled to an 80, 40 or 20% smaller value, or to a 40, 80 or 120% higher value. The range of resulting manipulated contrast values covered the range of contrast values in the original natural images. For each contrast setting and image a new mask was generated. A set of such manipulated images consisted of 40 times six images and animal M1 was presented with three sets (720 images) and animals M2, M3, and M4 viewed two sets (480 images).

A last set of images was used in the third paradigm (Fig. 4A). This was obtained using the same masking principle, but by blending patches from one natural image into another natural image. For each Gaussian kernel, a natural image different from the one being manipulated was chosen at random, and a patch was selected from a random location within this image. This patch, rotated randomly, was then blended into the initial image. Similar as for the noise, the mean luminance and contrast were scaled to that of the original image. For this condition, two manipulations were constructed based on the same mask, one with the contrast of the original image and one with a 120% increased contrast. A set of these images consisted of 40 times two images and animal M1 viewed three sets (240 images), animal M2 viewed two sets (160 images) and animals M3 and M5 each viewed one set (80 images).

### 2.2. Measurements of eye movements

Experiments were performed using five adult male rhesus monkeys (*Macaca mulatta*) that usually participate in other experiments at the Max Planck Institute for Biological Cybernetics. All procedures were approved by the local authorities (Regierungspraesidium Tübingen) and were in full compliance with applicable guidelines (EUVD 86/609/EEC) for the use of laboratory animals. During the experiments, the animals were head-fixed and sitting in a primate chair in a darkened booth. The stimuli were presented on a 21 in gamma-corrected monitor at a distance

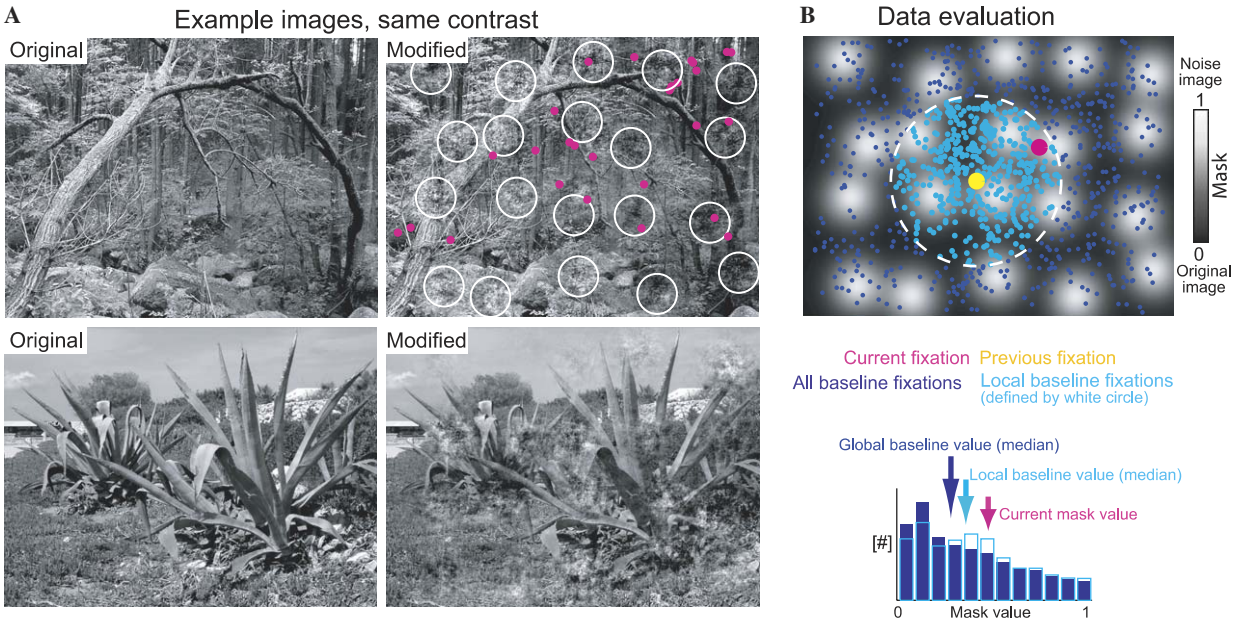


Fig. 1. Construction of image manipulations, example stimuli (first paradigm) and data evaluation. (A) The first paradigm used natural images (left) and manipulated images (right). The manipulations (white circles) consisted of locally blending a patch of pink noise into the image, such that the blended patch maintained the mean luminance and contrast of the original image at this location. Magenta dots show example fixations from one image presentation. (B) Fixations were characterized using the value of the mask at their location (actual value). A baseline value was computed, to estimate the mask value to be expected by chance. Several baselines were used (see Section 2). For one baseline (global baseline), 1000 random fixations were used, which the same animal made on any other image (blue dots). The median of the mask at these random fixations defined the baseline for the current fixation (see histogram). For a second baseline (local baseline), 1000 random fixations were used, which the same animal made on any other image, but with the additional constraint that these be within a region around the previous fixation (light blue dots within white circle). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

of 97 cm from the animal and covered a visual field of $24 \times 18$ deg. Images were presented for 8 s separated by a blank screen of variable duration (2–4 s) during which random juice rewards were delivered to the animal (probability of 0.7). Images of the same conditions were presented in random order, and images of different conditions were presented in separate blocks. During a single session only 30 or 45 images were presented to preserve the natural interest of the animals and prevent biases introduced by a forced watching of images. Sessions were repeated on a daily basis during several weeks. Eye movements were measured using the scleral search coil technique (Judge, Richmond, & Chu, 1980) and sampled at 200 Hz (CNC Engineering, Connecticut, USA). Before actual data acquisition, the eye coil was calibrated using a grid of nine fixation points (covering a range of $\pm 8$ deg), and each point was fixated six times within a window of 0.8 deg. Based on these fixations, the system was calibrated with an average spatial accuracy (r.m.s. error) of $0.51 \text{ deg} \pm 0.27 \text{ deg}$ (mean $\pm$ SD across animals).

### 2.3. Data analysis

From the eye movement recordings we extracted fixations as periods with only little eye movement following a fast saccadic eye movement. Saccades were defined by thresholding the velocity at 30 deg/s, and a fixation was defined as the period after a saccade in which the eye moved less then 0.3 deg during at least 100 ms. After watching the image for a while, it could be that the animal's eyes wandered to a region outside the monitor. This could happen as the animals were not imposed with a specific task but were watching these images at their own choosing. For analysis we used only those image presentations where at least six consecutive fixations were made on the image. Hence, our analysis of the time courses is limited to the first six fixations on an image. Analyzing longer sequences of fixations would require more subsequent fixations on the image and hence would reduce the number of trials available for analysis. The number six was chosen as it yields a good compromise between the number of valid trials and the length of the fixation sequence. Previous studies reporting an effect of time on the fixation placement found the prominent effects during the first four to six fixations (Henderson & Hollingworth, 1999; Parkhurst et al., 2002; Tatler et al., 2005).

The target of fixations was quantified using the mask underlying the image manipulation (Fig. 1B). The value of this mask was extracted at each fixation. It could range between 0 (unmodified) and 1 (completely modified). For statistical analysis, it is necessary to establish a baseline representing the null-hypothesis of no effect; hence a random placement of fixations. We tested three different baselines. The first baseline assumes a globally uniform placement of fixations and uses the median value of the entire mask as an estimate of what value could occur by chance. This baseline, however, does not take into account possible biases in the fixation placement of individual subjects. Such a bias could for example be a preference to look at the center of the monitor, or to one particular side, as has been reported frequently in previous studies (Henderson & Hollingworth, 1999; Parkhurst et al., 2002; Tatler et al., 2005). The second baseline accounts for this bias and uses 1000 randomly chosen fixations, which the same subject made on any other image. The mask of the present image of investigation is then sampled using these random fixations, and the median of this sample serves as a final baseline estimate (see Fig. 1B, blue dots and histogram). While this baseline takes individual biases into account, there is one problem to it. Under most natural circumstances, the planning and execution of eye movements depends on local information of the image and the amplitude of saccadic eye movements is highly skewed towards small displacements (see Fig. 2E and Itti, 2005b; Itti & Koch, 2001; Parkhurst et al., 2002; Wolfe, O'Neill, & Bennett, 1998). In other words, if the current fixation is near the left edge of the monitor, the next fixation is likely to be on the left side, even if the subject would show a general bias to the right. The third baseline takes this 'locality' of saccadic eye movements into account and uses 1000 randomly chosen fixations of the same subject, but with the constraint that these are within a region of interest around the previous fixation (12 deg circular window,

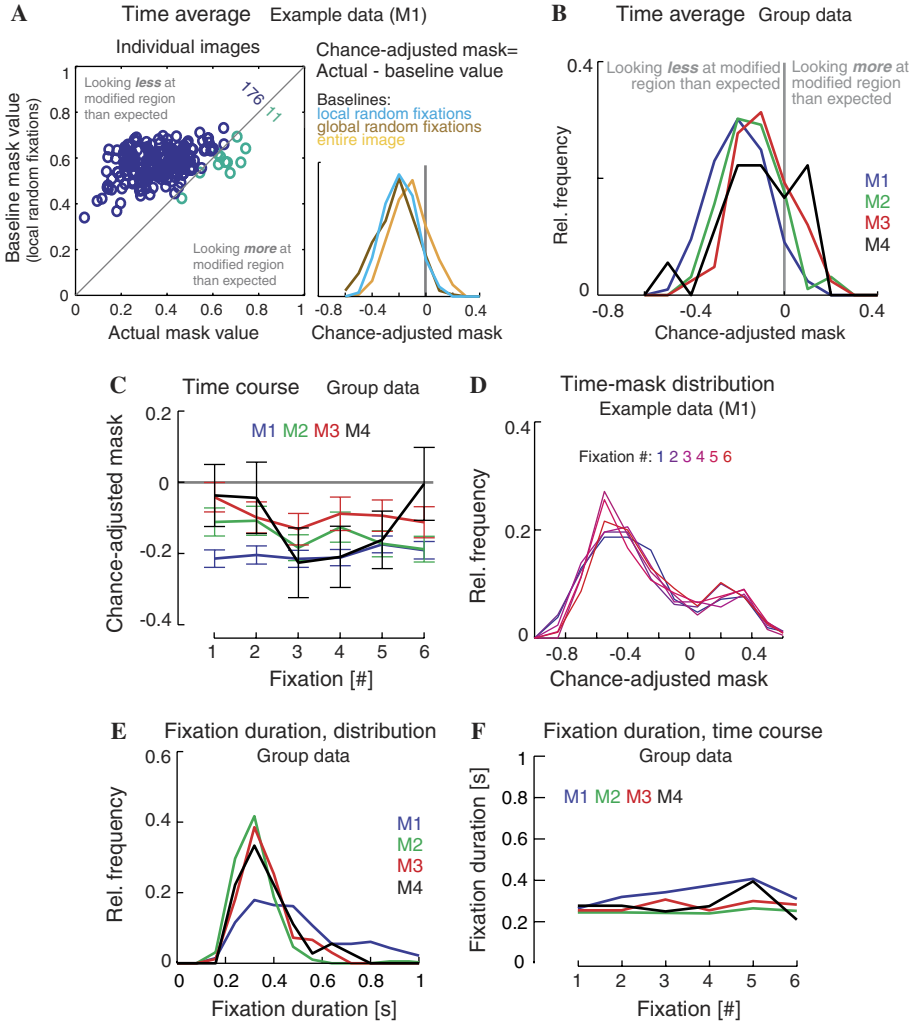


Fig. 2. Results from the first paradigm (noise patches in natural images). (A) Scatter plot (left panel) of the average actual and baseline (local random fixations, third baseline definition) mask value for individual image presentations to animal M1. For the dark circles, the baseline value was larger than the actual value, indicating a fixation preference for the natural, unmodified image regions. The histograms (right) show the same data in form of chance-adjusted mask values (actual minus baseline value) for all three different baseline definitions (see Section 2 for details). The results were consistent across baselines. (B) Group data from all animals. The histograms show the distribution of chance-adjusted mask values, averaged across individual image presentations (here and in the following always the 'local random fixations' baseline is used). Negative values indicate a preference for the not manipulated image regions. (C) Time courses of the mask value for individual animals (mean and SEM). (D) Distribution of the mask value at individual fixations (first, second, etc.) across image presentations for one animal. (E) Distribution of fixation duration for individual subjects across image presentations. (F) Time course of the fixation duration. The diagram shows the median value for each animal (solid lines).

see Fig. 1B, white circle). The median mask value at these random fixations then serves as baseline value. This baseline has been successfully used in several recent studies (Parkhurst et al., 2002; Tatler et al., 2005).

To determine whether any results depend on the choice of the baseline, we repeated all analyses for all three baselines. We did not find any difference in the qualitative results, nor was the significance or insignificance of any finding changed by the type of baseline. To illustrate this, Fig. 2A analyses the data from one example subject and displays the result using all three baseline methods (histograms in the right panel). For each baseline the histogram was shifted significantly towards negative values (Wilcoxon tests, at least $p < 10^{-5}$, for all three cases). As a further example, the ANOVA test across subjects in Fig. 2B was significant for all three baselines (effect of actual versus baseline mask values: $p = 0.0014$, $p \approx 0$, and $p \approx 0$ for the first, second, and third baseline, respectively). Hence, for the rest of the manuscript we used only the third baseline method (accounting for bias and saccade size) and all results are displayed as '*chance-adjusted*' mask value, which is defined as the difference between the actual mask value minus the baseline value.

For statistical analysis, we proceeded as follows. First, we computed the time averaged data by averaging across all fixations on a trial. These average mask values for all tested images were then subjected to a group analysis using an ANOVA for repeated measures, with animals as repeats and condition as factor. While the original mask values are distributed between 0 and 1, the chance-adjusted mask values are distributed around zero (e.g., see histograms in Fig. 2B) and allow the use of parametric statistics. In addition, we confirmed the results from the ANOVA for individual subjects using non-parametric tests such as the Wilcoxon rank-sum test (comparing actual vs. baseline mask values). The data for individual time points (fixation number) was analyzed using an ANOVA with time as a factor. To confirm negative findings from the ANOVA, a second non-parametric analysis was performed. For each time point and subject, a frequency histogram of the mask values was computed (see e.g., Fig. 2D). These histograms were then compared across subjects and time using a non-parametric ANOVA (Scheirer–Hare extension of the Kruskal–Wallis test, with subjects as repeats and time and histogram bins as factors (Sokal & Rohlf, 1995). In addition, for each subject we compared

the frequency histogram of the first time point to the histograms of all later time points using $\chi^2$ tests. For this comparison, effects were termed significant if they reached a $p$ value of 0.01 corrected for multiple comparisons (Bonferoni correction).

In addition to this quantification of where fixations were placed, we also computed the duration of individual fixations (see e.g., Fig. 2E). This analysis has been shown in previous studies to reveal effects of eye movement strategies (Henderson & Hollingworth, 1999). The fixation duration was similarly analyzed as the mask value, first as time averaged data and then using time as factor.

## 3. Results

In an initial experiment we established that monkeys would indeed study natural images, even when presented outside a specific task. In the same experiment we also presented pink-noise images, to compare general fixation properties on natural images and images devoid of clearly recognizable structures. On average, the animals seemed to be more interested in the natural images (Table 1). Both the time spent looking at the image was longer, and the number of fixations per second was higher for the natural images. Thus, the pattern of eye movements on natural images seemed to be more elaborate, indicating that the animals perceived these two classes of stimuli as different. Nevertheless, they spent considerable time looking at the noise images, demonstrating that such a free-viewing paradigm can yield usable data.

### 3.1. Fixations specifically avoid local noise patches

To quantify the influence of local information content on the fixation pattern, we locally introduced noise patches into natural scenes (Fig. 1A). At 20 randomly distributed locations, a pink-noise patch was blended into the original image. This modification preserves the local stimulus structure as defined by spatial frequency contrast, as the pink noise has the same amplitude spectrum as the original image. Yet, within the modified patch all object-like features were removed, as the phase spectrum of the noise was random. Hence, the global gist of the image was preserved, while locally the modified regions did not convey any information about the global image. Whether monkeys preferentially fixated normal or modified image regions was quantified using the mask which defines the image manipulation (Fig. 1B). For statistical analysis, the mask values at actually fixated locations were compared to baseline estimates, representing the null-hypothesis of no effect (Fig. 2A). We tested three different baselines as explained in more detail in the Section 2 and in Fig. 1B. However, the choice of baseline did not affect the main finding and significance of results (see Fig. 2A and Section 2).

Overall, the fixation pattern reflected the modifications introduced into the images. The data from the example animal reveal that, across images, the animal was looking less at the modified image patches than expected from the baseline (Fig. 2A, left panel): The average mask value at the actually fixated locations was smaller than the baseline value for 176 of 187 image presentations; only 11 trials showed the opposite effect (Wilcoxon test, $p = 0$). A similar result was found for all four animals tested, as shown by the histograms in Fig. 2B. Clearly, all histograms are shifted towards negative values, demonstrating that fixations preferentially target the unmodified image regions. Statistically, this result was confirmed by an ANOVA showing a highly significant effect of actual versus baseline mask values ($F(1,804) = 153$, $p \approx 0$), a weak effect of subjects ($F(3,804) = 3.4$, $p < 0.05$), and a significant interaction ($F(3,804) = 9.6$, $p < 10^{-5}$). We confirmed the strong effect of actual versus baseline fixation for each individual subject using a non-parametric test (Wilcoxon test, $p < 10^{-6}$ for animals M1–M3, and $p < 0.05$ for animal M4). Together, these results demonstrate unequivocally that, on average, fixations preferentially target the unmodified natural parts of the image and spare the uninformative noise patches.

This preference for natural image parts over uninformative noise patches was unchanged over time. In a first analysis, possible effects of fixation number were quantified using the time-course of the mask value for individual animals (Fig. 2C). No animal displayed a clear trend over time. This null result was confirmed by an ANOVA showing no effect of time ($F(5,2258) = 1.3$, $p = 0.22$), an effect of subjects ($F(3,2258) = 10.1$, $p < 10^{-5}$) and no interaction ($F(15,2258) = 0.6$, $p = 0.84$). To further substantiate this negative result, we used a non-parametric approach and investigated the distribution of the mask values along time for all images. This was done by computing the histogram

Table 1
General properties of eye movements

| | % time looking at image | # fixations/s | Sacc. ampl. | Fix. dur. |
|---|---|---|---|---|
| Natural images | 63 ± 11 | 2.5 ± 0.14 | 6.3 ± 0.8 | 383 ± 89 |
| Noise images | 44 ± 10 | 2.2 ± 0.17 | 8.8 ± 2.0 | 441 ± 90 |
| Noise same cont. | 54 ± 12 | 2.3 ± 0.16 | 8.1 ± 1.8 | 420 ± 82 |
| Noise modif. cont. | 49 ± 8 | 2.4 ± 0.17 | 8.9 ± 1.0 | 408 ± 84 |
| Natural same cont. | 57 ± 11 | 2.4 ± 0.32 | 6.0 ± 0.8 | 358 ± 86 |
| Natural incr. cont. | 58 ± 9 | 2.5 ± 0.35 | 5.7 ± 0.9 | 339 ± 81 |

The table displays the percentage of time the animal spent looking at the image, the number of fixations per second, the amplitude of saccadic eye movements (in degrees) as well as the duration of fixations (in ms) on average across animals (mean ± SD). The different conditions are natural images, noise images, natural images with blended noise patches (same and modified contrast conditions) and natural images with blended delusive natural patches (same and increased contrast conditions).

of mask values for each fixation number across images. Examples are displayed for one animal in Fig. 2D. To test an effect of time, we used a non-parametric ANOVA (Kruskal–Wallis test, see 2), which revealed no effect of time ($\chi^2 = 1.2$, $p = 0.94$), no effect of subjects ($\chi^2 = 3.4$, $p = 0.33$), and no interaction ($\chi^2 = 3.1$, $p = 0.91$). In addition, $\chi^2$ tests were performed for individual animals comparing the histogram of the first fixation to that of any later fixation. No comparison reached significance (at $p < 0.01$, including Bonferoni correction for multiple comparisons). Together, these results led us to conclude that there is no significant effect of fixation number on the distribution of fixations with respect to natural and manipulated parts in the image.

In our initial experiments we found that the average duration of fixation was different on completely natural and noise images (Table 1). Hence, if there was a temporal effect on fixation placement, this could manifest as different fixation durations for initial and later fixations. To test this, we investigated the fixation duration in this paradigm. On average, there was some variability between subjects concerning the typical duration of fixations (Fig. 2E). Especially, animal M1 tended to make longer fixations (median value 310 ms) compared to the other animals (250, 275, and 245 ms). However, there was no clear effect of fixation number on the distribution of fixation duration (Fig. 2F). This negative finding was confirmed by an ANOVA (no effect of time $F(5,2258) = 0.6$, $p = 0.67$; significant effect of subject $F(3,2258) = 51$, $p \approx 0$; and no interaction $F(15,2258) = 1.0$, $p = 0.41$). Hence, we can conclude that both the fixation locations, as well as their duration did not exhibit any systematic change during inspection of a visual scene.

### 3.2. Contrast manipulations can compensate for missing informative content

The first experiment demonstrated that informative scene content influences the fixation pattern. In the second paradigm, we investigated the interaction of physical scene structure, as defined by the amplitude of the Fourier spectrum, and informative scene content. One property derived from the amplitude spectrum, luminance–contrast, has been shown to play an important role in determining fixated locations: Several studies showed that luminance–contrast is especially high at fixated locations (Einhäuser & König, 2003; Krieger et al., 2000; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). To probe whether changes in luminance–contrast interact with fixation placement, we modified the contrast of the blended pink-noise patches (Fig. 3A). The range of these contrast modifications was chosen similar to previous studies (Einhäuser & König, 2003) and fully covers the range of contrast values occurring naturally in these images.

Both enhancing and decreasing the contrast of the noise patches attracted fixations towards them. Fig. 3B exemplifies this for one subject and displays the mask value for all contrast conditions, including the unmodified contrast condition (0% modification). For all modifications, the mask value was significantly larger than in the unmodified condition (Wilcoxon tests, $p < 0.001$ each comparison). Manipulating the contrast thus led to a shift of fixations towards the modified patches. This shift either means that more fixations specifically targeted the center of the manipulated regions, or that the fixations in general fell closer to these regions (see below). For the contrast increments (+40, +80, and +120%), the resulting distribution of chance-adjusted mask values was not significantly different from zero (Wilcoxon tests, $p > 0.05$, for all three comparisons). Hence, the increased contrast compensated the effect of the missing scene content and led to a pattern of fixations not distinguishable from baseline. This effect of contrast manipulation was replicable in all four animals (Fig. 3B, lower panel) and confirmed by an ANOVA (significant effects of condition $F(6,1994) = 35$, $p = 0$, and subjects $F(3,1994) = 38$, $p \approx 0$; and an interaction $F(18,1994) = 4.1$, $p < 10^{-6}$). For individual subjects we confirmed that each contrast modification was significantly different from the unmodified condition (Wilcoxon tests, at least $p < 0.05$ for all comparisons, Bonferoni corrected). Hence, we conclude that altering the local image structure by introducing conspicuous contrast modifications can compensate for the missing informative content in the noise patches and attracts fixations.

As stated above, the trend of the chance-adjusted mask value towards zero implies that either the fraction of the fixations right in the center of the modified patches increased, or that overall most fixations were placed closer to the modified regions. To tease these two possibilities apart, we performed an addition analysis: A $\chi^2$ test was used to compare the frequency of fixations at image regions with a mask value higher than 0.8, hence being close to the peak of the modification. The test compared the frequency observed in same contrast condition with that observed in the contrast-modification paradigm (+120%) and resulted in a highly significant difference ($\chi^2 = 102$, $p \approx 0$). Thus, the modifications in luminance–contrast specifically attracted fixations towards the center of the uninformative noise patches.

We did not find any evidence for an effect of time in the contrast modified conditions. Fig. 3C displays the time-courses of each animal averaged across conditions with contrast increases. There was no consistent effect observable, and this was confirmed by the ANOVA (no effect of time $F(5,3888) = 1.2$, $p = 0.30$; no effect of subjects $F(3,3888) = 1.6$, $p = 0.18$; and no interaction $F(15,3888) = 1.4$, $p = 0.09$). In addition we compared the histograms of the mask values at different fixation numbers (Fig. 3D). The non-parametric ANOVA demonstrated no effect of time ($\chi^2 = 1.6$, $p = 0.9$), an effect of subjects ($\chi^2 = 20.8$, $p < 0.001$), but no interaction ($\chi^2 = 4.9$, $p = 0.85$), and the $\chi^2$ comparison of individual time points did not reveal a significant effect in any animal. For those conditions with contrast decreases, the same analysis was

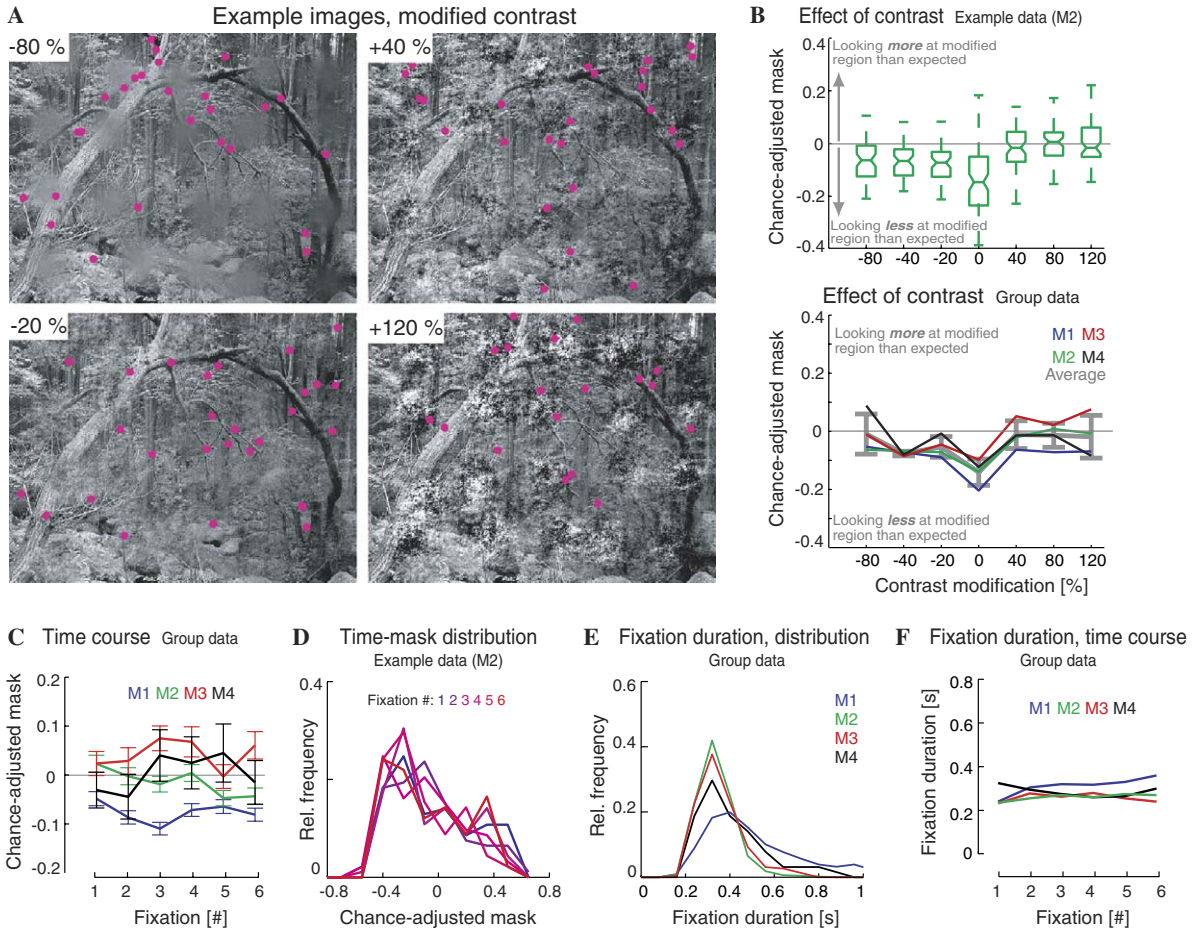


Fig. 3. Results from the second paradigm (noise patches with manipulated contrast in natural images). (A) Example images with different contrast manipulations. The contrast of the noise patches was scaled at different levels with respect to the original contrast (insets in top left corners). Magenta dots indicate fixations during one presentation. (B) Average chance-adjusted mask value at fixations. Upper panel: data for individual images and one animal (boxes indicate median and upper and lower quartiles, lines indicate the range of data). Zero contrast modification reproduces the data from the condition in Fig. 2. Lower panel: group data; median values for individual animals, as well as the mean and SD across animals (gray line). (C) Time courses of the mask value for individual animals (mean and SEM). (D) Distribution of the mask value at individual fixations (first, second, etc.) across image presentations for one animal. (E) Distribution of fixation duration for individual subjects across image presentations. (F) Time course of the fixation duration. The diagram shows the median value for each animal (solid lines). The data in (C–F) were averaged across all paradigms with contrast increases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

applied (not shown in the figures): The ANOVA demonstrated no effect of time ($F(5,4548) = 1.0$, $p = 0.39$), no effect of subjects ($F(3,4548) = 0.5$, $p = 0.66$) and no interaction ($F(15,4548) = 1.23$, $p = 0.12$). Also, the non-parametric analysis of the mask histograms did not show any significant effect (no effect of time $\chi^2 = 1.7$, $p = 0.88$; a weak effect of subjects $\chi^2 = 14$, $p < 0.01$; no interaction $\chi^2 = 3.4$, $p = 0.9$). Together these results show that there is no clear and significant effect of time on the fixation placement in either the contrast increase or decrease paradigms.

Also in this paradigm, there was no effect on the duration of fixations. Fig. 3E displays the distribution of fixation durations for the individual animals. We found no difference between the contrast increase and contrast decrease paradigms and the results are in good agreement with those found without contrast modification (c.f. Fig. 2E). Plotting the distribution of fixation duration as a function of fixation number did not show a clear effect (Fig. 3F, for the contrast increases), as confirmed by the ANOVA (no effect of time $F(5,3888) = 0.7$, $p = 0.61$; a significant effect of subjects $F(3,3888) = 11$, $p < 10^{-6}$; no interaction $F(15,3888) = 1.5$, $p = 0.08$). The same finding holds for the contrast decreases ($F(5,4548) = 0.2$, $p = 0.96$; $F(3,4548) = 4.6$, $p < 0.01$; and $F(15,4548) = 0.5$ $p = 0.90$ respectively). Thus, we can also conclude for this paradigm that fixation placement and duration do not show significant changes throughout the time course of picture viewing.

### 3.3. Uninformative natural patches are more attractive than noise

The two paradigms above used pink-noise patches that were locally blended into natural images. The results showed that monkeys preferentially fixate the natural

image part and inspect the noise patches only rarely. Hence, the eye movements preferentially targeted the more informative natural part of the image and spared those parts devoid of clearly recognizable structures. It might be that this image manipulation resulted in a bias that was too strong to observe clear changes of fixation strategy along time. For example, even in the contrast modified paradigm, fixations never occurred significantly more frequently on the modified patches than expected. As a consequence, we designed a third paradigm and used patches from another natural image instead of pink-noise: We blended patches of one natural image into another image, once preserving the contrast in the image and once increasing the contrast (Fig. 4A). Hence, the modifications again preserved local luminance and contrast, but—in contrast to the above paradigms—the blended patch was not devoid of natural structures, but had a natural like amplitude and phase spectrum. Yet, also in this paradigm, the manipulated patches did not convey information about the global image content. Both the same contrast and the increased contrast blending were constructed using the same mask to allow a direct comparison of the two conditions.

When introducing patches from one natural image into a different image, their effect depended on contrast, as in the previous paradigm. Fig. 4B displays the distribution of mask values for the two conditions and four animals. Across animals, blending same contrast and increased contrast patches led to different effects on the fixation placement. An ANOVA showed a strong effect of condition ($F(1,768) = 77$, $p \approx 0$), no effect of subjects ($F(3,768) = 2.2$, $p = 0.08$) but an interaction ($F(3,768) = 6.1$, $p < 0.001$). The effect of condition was confirmed for each subject by a Wilcoxon test ($p < 0.01$, all comparisons). For the same contrast condition, three of four subjects showed a significant trend towards the unmodified parts of the image—that is away from the modifications (see $p$ values in Fig. 4B). This trend is similar to what was found above with the pink-noise patches. In both cases fixations spared the manipulated patches, which are inconsistent with the global gist of a scene. For the increased contrast condition, fixations were specifically attracted towards the

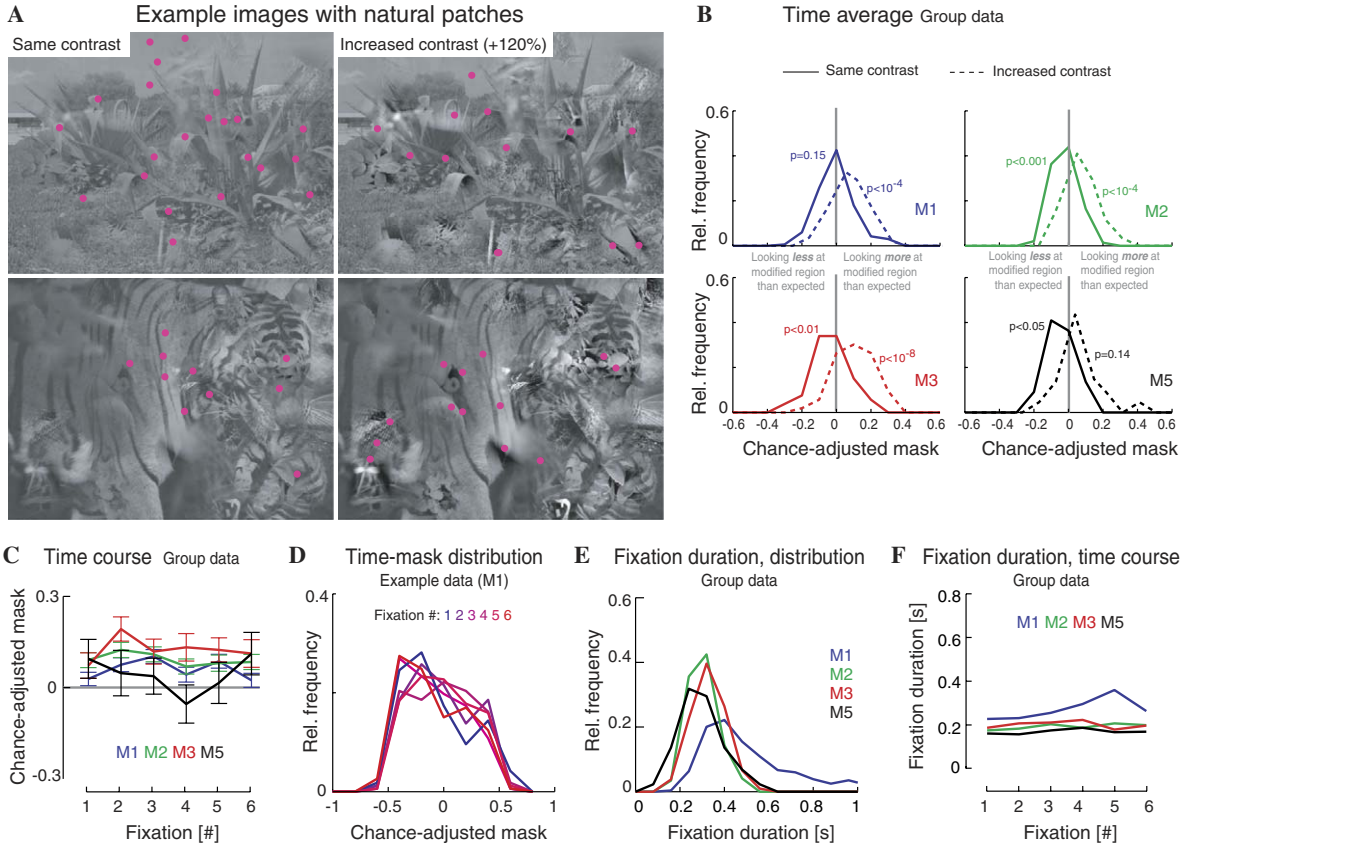


Fig. 4. Results from the third paradigm (natural patches in natural images). (A) Example images. In this manipulation, a patch from one natural image was blended into a different image, once by preserving the contrast of the target image (left column) and once by enhancing the contrast by +120% (right column). (B) Group data showing the average chance-adjusted mask value for both conditions and individual subjects. $p$ values at individual histograms indicate the statistical significance for the median being different from zero (Wilcoxon test). (C) Time courses of the mask value for individual animals (mean and SEM). (D) Distribution of the mask value at individual fixations (first, second, etc.) across image presentations for one animal. (E) Distribution of fixation duration for individual subjects across image presentations. (F) Time course of the fixation duration. The diagram shows the median value for each animal (solid lines). The data in (C–F) are from the contrast increase paradigm.

modification, and this effect was significant in three out of four animals (see $p$ values in Fig. 4B). Altogether, these results confirm the above finding that increasing the contrast of local patches attracts more fixations towards these.

Similar to the first two paradigms, there was no consistent effect of fixation number in either of these paradigms (Fig. 4C, for the contrast increase paradigm). The ANOVA of the time-courses did not reveal any significant effect of time (same contrast condition: no effect of time $F(5,2248) = 0.45$, $p = 0.81$; no effect of subjects $F(3,2248) = 0.86$, $p = 0.45$; and no interaction $F(15,2248) = 0.86$, $p = 0.6$; increased contrast condition: no effect of time $F(5,2273) = 1.0$, $p = 0.42$; significant effect of subjects $F(3,2273) = 4.9$, $p < 0.01$; and no interaction $F(15,2273) = 0.75$, $p = 0.73$). The same result was obtained from the non-parametric analysis of the mask histograms (same contrast condition: no effect of time $\chi^2 = 0.15$, $p = 0.99$; no effect of subjects $\chi^2 = 0.79$, $p = 0.85$; and no interaction $\chi^2 = 0.92$, $p = 0.94$; increased contrast condition: no effect of time $\chi^2 = 0.12$, $p = 0.99$; no effect of subjects $\chi^2 = 0.78$, $p = 0.85$; and no interaction $\chi^2 = 0.13$, $p = 0.99$). Last, this was also confirmed by the $\chi^2$ comparisons of individual time points (data not shown).

The analysis of fixation duration was also consistent with the above results and did not reveal any effect of time (Figs. 4E and F). For both conditions the ANOVA showed no effect of time (same contrast: no effect of time $F(5,2248) = 1.7$, $p = 0.11$; significant effect of subjects $F(3,2248) = 51$, $p \approx 0$; no interaction $F(15,2248) = 1.44$, $p = 0.11$; increased contrast: no effect of time $F(5,2273) = 1.0$, $p = 0.37$ significant effect of subjects $F(3,2273) = 79$, $p \approx 0$; no interaction $F(15,2273) = 1.6$, $p = 0.052$). Hence, neither fixation placement nor fixation duration show a clear trend during inspection of an image.

## 4. Discussion

We analyzed the fixations of macaque monkeys that freely viewed natural and manipulated natural images without a particular task. The manipulations were designed to dissociate stimulus structure (luminance–contrast) and local information content to quantify the weighting of these for determining fixations. We found a strong bias away from regions with low information content within natural scenes. Increasing the contrast of the uninformative parts, however, attracted more fixations towards these and could compensate for missing content in the manipulated regions. Further, analyzing sequences of consecutive fixations, we could not find evidence for an effect of time on how stimulus manipulations affected fixation placement. Both initial and later fixations were similarly affected by the changes in local image structure or the informative image content.

Our results demonstrate that fixations specifically spare uninformative parts in natural scenes. This is in good agreement with previous findings that fixations in complex scenes target regions that are informative as rated by other human observers (Antes, 1974; Mackworth & Morandi, 1967), are informative for a given task (De Graef et al., 1990; Yarbus, 1967) or are informative based on their relation to the global gist of a scene (Henderson et al., 1999; Loftus & Mackworth, 1978). However, the definition of which parts in a scene are highly informative might well depend on a particular task imposed or might depend on subjective biases. In the present case, informative image content was manipulated by either randomizing the higher order structure of the local region (pink-noise), or by introducing patches from one natural image into another image. In both cases were the resulting patches inconsistent with the global image. Our results extend the previous findings, as we demonstrate a general bias in the way naïve monkeys freely view natural scenes. Instead of inspecting the manipulated patches, which are inconsistent with the global image, the fixations specifically targeted the unmodified parts of the image, which are informative about the image. Further, these results suggest that a peripheral analysis of the image plays an important role during eye movement planning. Such a peripheral analysis can guide eye movements specifically to highly salient or informative regions as demonstrated in previous studies (De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978) and significantly improves the performance of saliency map models in predicting the fixation pattern of human subjects in natural scenes (Itti, 2005b; Peters, Iyer, Itti, & Koch, 2005). Further, peripheral analysis operates very fast and is sufficient to recognize the gist of a scene (Rousselet, Joubert, & Fabre-Thorpe, 2005) and the quality of peripheral vision has clear influences on eye movement patterns (Loschky, McConkie, Yang, & Miller, 2005). In the present study, peripheral analysis must have been used to specifically avoid the noise patches when planning eye movements to new fixations points.

Increasing the saliency of local visual features attracted fixations towards the modified locations. Changes of luminance–contrast drove fixations to the noise patches despite a general bias away from these uninformative regions, and, in the extreme cases, such contrast enhancements canceled the repulsive effect of missing scene content. This demonstrates that contrast can compensate for missing information content. Such a strong effect of image contrast on fixations fits with previous findings. Several studies reported a significant correlation of contrast and human fixations in natural scenes (Krieger et al., 2000; Mannan et al., 1996; Parkhurst & Niebur, 2003; Parkhurst & Niebur, 2004; Reinagel & Zador, 1999; Tatler et al., 2005). Extending this, a recent study demonstrated that contrast manipulations indeed attract human fixations (Einhäuser & König, 2003). Einhaeuser and colleagues reported that both strong de- and increases of contrast, either below −40% or above 100%, had a significant attractive effect on fixations. Our results differ from this, as we found a significant effect of contrast even for the smallest modifications (−20%, +40%), which attracted fixations towards uninformative noise patches. A reason for this difference might be that

humans and macaque monkeys have different sensitivity to contrast modifications. Indeed, recent results from Einhaeuser and colleagues suggest that monkeys are much more susceptible to contrast modifications than humans (Einhäuser et al., 2006).

Several previous studies suggested that eye movements are guided both by spatial image structure and by image content, with the former being more prominent for the initial fixations and the latter guiding subsequent fixations (De Graef et al., 1990; Henderson et al., 1999; Itti, 2005a; Parkhurst et al., 2002) and see (Henderson & Hollingworth, 1999) for a review. These ideas were recently summarized in a model by assuming that the balance between bottom-up influence and top-down control of saccadic targets changes over time (Tatler et al., 2005). The present results however clearly demonstrate that such an effect does not persist during naïve free viewing of natural scenes. Hence, if such a trend is observed it could well result from a particular task imposed on the subject or an implicit strategy which the subject might be following.

The present study analyzed only the first six fixations on each image. The reason for this restriction is that the subjects were not imposed a particular task and hence not forced to keep their eyes within the image. Therefore, a trade off was necessary between, first, the number of trials with more than the required number of subsequent fixations within the image and, second, the number of required subsequent fixations (c.f. 2). However, the previous studies that found an effect of fixation number clearly showed that such effects should occur within the first four to six fixations (Henderson & Hollingworth, 1999; Parkhurst et al., 2002; Tatler et al., 2005), and thus clearly within the limit of the present analysis.

Fixation patterns are often very different between observers (Einhäuser et al., 2006; Henderson & Hollingworth, 1999; Tatler et al., 2005; Yarbus, 1967). To explain this variability, a recently proposed model suggests that human scan paths can be understood based on a *random selection with distance weighting* model (Melcher & Kowler, 2001). This model proposes that fixations target random locations that are only constrained based on the distance from the previous fixation. However, this model does not incorporate effects induced by scene content or scene structure and hence cannot explain the present findings. For example, this model would neither predict that fixations are specifically attracted towards the noise patches if the contrast of these is strongly increased nor that this effect is even more pronounced if natural patches instead of noise are used. Instead, we argue that the variability of fixations results from *individual strategic differences* between observers, but also between trials and tasks imposed. While the bottom-up influence of salient scene structure is constant over time, the influence of information content is integrated with the momentary strategy for exploring a scene and hence differs between individuals and possibly also between repeated presentations of the same image. This hypothesis is consistent both with the classical observations by Yarbus (Yarbus, 1967), as well as with recent results (Einhäuser et al., 2006; Itti, 2005a; Itti, 2005b; Tatler et al., 2005).

## Acknowledgments

This work was supported by the Max Planck Society and by the DFG (to CK, KA 2661/1). We would like to thank Shi-Pi Ku, Stefanie Liebe and Joost Maier for their help with collecting data.

## References

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103*, 62–70.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*, 317–329.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*, 1089–1097.

Einhäuser, W., Kruse, W., Hoffmann, K., & König, P. (2006). Differences in monkey and human overt attention under natural conditions. *Vision Research, 46*(8–9), 1194–1209.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 210–228.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology, 25*, 201–208.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology–General, 127*, 398–415.

Itti, E. (2005a). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*, 1093–1123.

Itti, E. (2005b). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, In Press.

Itti, L., & Koch, C. (1998). A model of Saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Learning, 22*, 1254–1259.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Judge, S. J., Richmond, B. J., & Chu, F. C. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research, 20*, 535–538.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13*, 201–214.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology–Human Perception and Performance, 4*, 565–572.

Loschky, L. C., McConkie, G. W., Yang, J., & Miller, M. E. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition, 12*, 1057–1092.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2*, 522–547.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10*, 165–188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11*, 157–178.

Melcher, D., & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research, 41*, 3597–3611.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*, 125–154.

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19*, 783–789.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*, 2397–2416.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*, 341–350.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get the gist of real-world natural scenes? *Visual Cognition, 12*, 852–877.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry*. New York: W.H. Freeman and Company.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

Wolfe, J. M., O'Neill, P., & Bennett, S. C. (1998). Why are there eccentricity effects in visual search. Visual and attentional hypotheses. *Perception Psychophysics, 60*, 140–156.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.